

A Comparative Analysis of Machine Learning and Artificial Intelligence-based Models for Diabetes Prediction

Jitendra Sheetlani¹, Ajay Vyas², Harsh Gupta³

^{1,2} Sri Satya Sai University of Technology and Medical Science, Sehore (MP.)

³IT Project Manager and Independent Researcher

ABSTRACT:- *Diabetes has become a significant global public health concern as the prevalence of non-communicable diseases continues to rise. Heart disease claims the lives of approximately 18 million people annually, with diabetes and high blood pressure emerging as primary contributing factors. The social, physical, and economic consequences of diabetes are substantial. Elevated blood sugar levels characterise this chronic condition and occur due to the body's inability to produce or properly respond to insulin. Data mining enables analysts to efficiently analyse extensive data sets to identify patterns and trends associated with diabetes. In recent years, machine learning (ML) methods have been utilised for diabetes prediction. Data mining involves extracting essential information and leveraging it to enhance dynamic effectiveness. Various AI techniques, including Support Vector Machine (SVM), Random Forest, Decision Tree (DT), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), and Naive Bayes (NB) classifiers, have been employed in diabetes prediction. This research focuses on analysing different machine-learning models for diabetes prediction. The paper is structured into sections: the first section discusses the various machine learning models and their distinct activation functions employed in this study. In contrast, the second section presents a comparative analysis of these models.*

Keywords: *Diabetes prediction, Machine learning, Random Forest, Support Vector Machines (SVM), Logistic regression (LR), Gradient boosting (GB), k-nearest neighbour (k-NN).*

INTRODUCTION

Diabetes is a widespread disease that affects people worldwide, with a higher prevalence observed in industrialised nations. Given its seriousness and impact, diabetes requires significant attention from medical professionals, patients, families, and society. The social, physical, and economic consequences of diabetes are substantial. It is a chronic condition characterised by elevated blood sugar or glucose levels. Diabetes arises from the body's inability to produce insulin or the failure of insulin to act on the cells effectively. The medical community has termed this phenomenon "x syndrome," as the exact cause of diabetes is still not fully understood. Historically, the focus of diabetes treatment has been on managing symptoms rather than addressing the underlying problem. Diabetes management is crucial in healthcare, attracting significant attention from researchers and medical practitioners. Using Machine Learning techniques allows researchers to analyse existing data and uncover patterns and trends related to diabetes, facilitating faster data processing and insights. Early detection of diabetes is a key concern in the healthcare sector. Several studies have utilised machine learning-based models and techniques to predict diabetes, often employing the PIMA Indian Diabetes Dataset (PIDD) accurately. Our research paper aims to analyse various machine learning models capable of accurately predicting diabetes.

LITERATURE REVIEW

Researchers have widely used the Pima Indian Diabetes Dataset (PIDD) to develop machine learning-based models for predicting diabetes, which describe

the characteristics of diabetes patients. This section reviews the literature that explores various machine-learning approaches for building an autonomous diabetes prediction system. Michie et al. provide an updated review of different classification approaches, comparing their performance on challenging datasets and discussing their applicability to real-world industrial problems. Constructing a classification procedure from known data with true classes is referred to as pattern recognition, discrimination, or supervised learning [1]. Jeatrakul et al. conducted a research paper comparing the performance of different neural network techniques for binary classification problems. They compared five types of neural networks, including Back Propagation Neural Networks (BPNN), Radial Basis Function Neural Networks (RBFNN), General Regression Neural Networks (GRNN), Probabilistic Neural Networks (PNN), and Complementary Neural Networks (CMTNN). The comparison was based on three benchmark datasets from the UCI machine learning repository [2]. Estebanez et al. utilised genetic programming-based data projections in their research. They presented a method based on genetic programming (GP) to automatically evolve projections, making data classification easier in the projected space. The approach involved evolving independent sub-trees, allowing for the construction of relevant attributes and potential dimensionality reduction or increase depending on the feasibility of classification in higher dimensional spaces [3]. Anjli Negi (2021) emphasises that diabetes is a metabolic disease characterised by impaired glucose utilisation, leading to consistently high blood glucose levels. Complications of diabetes include diabetic ketoacidosis, nonketotic hyperosmolar coma, heart disease, stroke, chronic renal failure, retinal damage, and foot ulcers. The global prevalence of diabetes is rapidly increasing, posing a significant public health concern. Early detection of diabetes plays a crucial role in reducing the risk of major complications and facilitating effective treatment [4]. A. M. Abdulazeez (2021) highlights the wide application of machine learning (ML) in computational work for algorithm

development and performance improvement. Learning from unbalanced datasets has been a significant challenge in machine learning, appearing in various applications such as computer security, Swarm Intelligence, remote sensing, and biomedicine [5].

COMPARATIVE ANALYSIS OF DIFFERENT MODELS

Researchers have extensively investigated the Pima Indian Diabetes dataset using various machine learning algorithms to predict diabetes. Table 1 displays the performance of different algorithms in terms of correct classification and miss classification (error rate) on the dataset. According to the table, some algorithms achieved the highest correct classification percentages. The “Logdisc” algorithm achieved a correct classification of 77.7% with a miss classification (error rate) of 22.3%. Similarly, the “DIPOL92” algorithm achieved a correct classification of 77.6% with a miss classification of 22.4%, while the “Discrim” algorithm achieved a correct classification of 77.5% with a miss classification of 22.5%.

On the other hand, certain algorithms exhibited higher miss classification percentages or error rates. For instance, the “K-NN” algorithm achieved a correct classification of 67.6% with a miss classification of 32.4%. Similarly, the “ALLOC80” algorithm achieved a correct classification of 69.9% with a miss classification of 30.1%. The average correct classification percentage across all algorithms was 73.8%, with an average miss classification of 26.2%. When selecting the most suitable algorithm for the task, it’s crucial to consider factors such as specific prediction requirements, dataset characteristics, and the trade-off between correct and miss classification rates.

Extensive research has been conducted to improve the order precision of predictions on the Pima Indian Diabetes dataset using artificial neural networks. In this regard, Jeatrakul and Wong examined the performance of various neural network architectures, including Back Propagation Neural Network (BPNN),

General Regression Neural Network (GRNN), Radial Basis Function Neural Network (RBFNN), Probabilistic Neural Network (PNN), and Complementary Neural Network (CNN), also known as Complementary Multi-task Neural Network (CMTNN).

Table 1: Pima Indian Diabetes Dataset Michie, Spiegelhalter, and Taylor Classification Results

S No	Algorithm	CC (%)	ER (%)
1	Discrim	77.5	22.5
2	Quadisc	73.8	26.2
3	Logdisc	77.7	22.3
4	SMART	76.8	23.2
5	ALLOC80	69.9	30.1
6	K-NN	67.6	32.4
7	CASTLE	74.2	25.8
8	CART	74.5	25.5
9	IndCART	72.9	27.1
10	NewID	71.1	28.9
11	AC2	72.4	27.6
12	Baytree	72.9	27.1
13	NaiveBay	73.8	26.2
14	CN2	71.1	28.9
15	C4.5	73	27
16	Itrule	75.5	24.5
17	Cal5	75	25
18	Kohonen	72.7	27.3
19	DIPOL92	77.6	22.4
20	Backprob	75.2	24.8
21	RBF	75.7	24.3
22	LVQ	72.8	27.2
23	Average	73.80	26.20
CC= Correct Classification, ER Error rate			

Table 2 presents the performance results of these architectures. The table provides the precision or accuracy achieved by each neural network architecture across multiple tests on the Pima Indian Diabetes dataset. Here is a summary of the results: Test 1: BPNN achieved a precision of 77.27%, GRNN - 74.68%, RBFNN - 79.22%, PNN - 74.68%, and

CMTNN - 77.92%. Test 2: BPNN achieved a precision of 76.62%, GRNN - 79.87%, RBFNN - 79.22%, PNN - 79.87%, and CMTNN - 76.62%. Test 3: BPNN achieved a precision of 70.13%, GRNN - 70.13%, RBFNN - 74.03%, PNN - 70.13%, and CMTNN - 72.08%. Test 4: BPNN achieved a precision of 85.71%, GRNN - 81.82%, RBFNN - 79.22%, PNN - 81.82%, and CMTNN - 83.77%. Test 5: BPNN achieved a precision of 75.97%, GRNN - 75.97%, RBFNN - 77.27%, PNN - 75.97%, and CMTNN - 75.32%. These findings offer valuable insights into the performance of different neural network architectures on the Pima Indian Diabetes dataset. The average precision ranges from 75.26% to 76.56% across all architectures.

Table 2: Results of Jeatrakul and Wong's classification of the Pima Indian Diabetes Dataset

TN	BPNN	GRNN	RBFNN	PNN	CMTNN
1	77.27	74.68	79.22	74.68	77.92
2	76.62	79.87	79.22	79.87	76.62
3	70.13	70.13	74.03	70.13	72.08
4	85.71	81.82	79.22	81.82	83.77
5	75.97	75.97	77.27	75.97	75.32
6	70.78	70.13	72.08	70.13	72.08
7	75.32	72.73	76.62	72.73	75.97
8	79.22	78.57	77.27	78.57	79.22
9	74.68	74.68	76.62	74.68	75.32
10	75.97	74.03	74.03	74.03	76.62
AVG	76.17	75.26	76.56	75.26	76.49
TN=Test Number and AVG=Average					

Estebanez, Alter, and Valls employed genetic programming-based data projections to analyse clustering tasks.

Table 3: Pima Indian Diabetes Dataset Classification by Estebanez, Alter, and Valls

SNo.	Algo	Accuracy
1	SVM	77.21
2	Simple Logistics	77.86
3	Multilayer Perceptron	76.69

They utilised the Pima Indian diabetes dataset and reduced the data dimensionality from 8 to 3. They employed Support Vector Machine (SVM), Simple Logistics, and Multilayer Perceptron algorithms for the classification task, utilising the Pima Indian Diabetes data. The results of their analysis are presented in Table 3.

The Multilayer Perceptron of the Fake Neural Network achieved a characterisation performance of 76.69 percent. Single Logistics, on the other hand, obtained the highest rating of 77.86 percent. In a study by Lena Kallin Westin, she investigated various preprocessing approaches to handle missing data in the Pima Indian Diabetes dataset. She developed a preprocessing perceptron for decision support using the diabetes dataset. The trained decision support network achieved an overall classification performance of 79 percent. In another research by Bylander, different classification approaches, including Naive Bayes, decision trees, and two types of ensemble methods, were employed on the Pima Indian Diabetes dataset. Table 4 presents the various classification approaches by Bylander and their corresponding classification performance.

Table 4: Pima Indian Diabetes Dataset Bylander Classification Performance

Sr. No.	Method	Accuracy
1	Belief Network(Laplace)	72.50%
2	Belief Network	72.30%
3	Decision Tree	72.00%
4	Naïve Bayes	71.50%

Misra and Dehuri made a Functional Link Artificial Neural Network for Classification Task in Data Mining. They stood apart its depiction execution from other AI calculations in their review Functional Link Artificial Neural Network for Classification Task in Data Mining. Their FLANN gathering execution was 78.13 percent, while their MLP demand execution was 75.2 percent. On the Pima Indian Diabetes dataset, Table 5 shows the strategy execution of different AI

calculations.KNN from case-based thinking can be utilised to recover comparative authentic models and eliminate exceptions, bringing about upgraded brain network arrangement execution.

Table 5: Performance of Misra and Dehuri Classification on Pima Indian Diabetes Dataset

Sr. No.	Model Name	Accuracy
1	NN	65.1
2	KNN	69.7
3	FSS	73.6
4	BSS	67.7
5	MFS1	68.5
6	MFS2	72.5
7	CART	74.5
8	C4.5	74.7
9	FID3.1	75.9
10	MLP	75.2
11	FLANN	78.13

CONCLUSION

we conducted a comparative analysis of machine learning and AI-based models for predicting diabetes with high accuracy. We employed several machine learning algorithms, such as Decision Tree (DT), K-Nearest Neighbor (KNN), and Logistic Regression (LR), on the PIMA Indian Diabetes Dataset (PIDD) to predict diabetes and evaluated their performance based on different parameters. The results obtained from these models were promising, showing good performance across various metrics. The early detection of diabetes is important in addressing the health challenges associated with the disease. By leveraging machine learning and AI techniques, we can effectively predict diabetes and enable timely intervention and treatment, thus mitigating the potential complications associated with the condition. Our study highlights the potential of machine learning and AI-based approaches in accurately predicting diabetes. Further research and advancements in this field can contribute to developing more robust and

efficient models for the early detection and management of diabetes, ultimately improving the overall healthcare outcomes for individuals affected by the disease.

REFERENCES

- [1]. Michie, D., Spiegelhalter, D. J., and Taylor, C. C. (1994). Machine Learning, Neural and Statistical Classification (Chapter 9, pp. 157-158). Pearson Education Limited.
- [2]. Jeatrakul, P., and Wong, W. K. (2009). Comparing the Performance of Different Neural Networks for Binary Classification Problems. In Proceedings of the Eighth International Symposium on Natural Language Processing (pp. 111-115).
- [3]. Estebanez, C., Aler, R., and Valls, M. (2005). Genetic Programming Based Data Projections for Classification Tasks. World Academy of Science, Engineering and Technology, 11(11), 56-61.
- [4]. Misra, B. B., and Dehuri, S. (2007). Functional Link Artificial Neural Network for Classification Task in Data Mining. Journal of Computer Science, 3(12), 948-955.
- [5]. Kavakiotis, I., et al. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104-116.
- [6]. Kavakiotis, I., et al. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104-116.
- [7]. Negi, A. (2021). A review of current advances in machine learning-based diabetes research.
- [8]. Abdulkareem, N. M., and Abdulazeez, A. M. (2021). Machine Learning Classification Based on Random Forest Algorithm: A Review. International Journal of Science and Business, 5(2), 128-142.
- [9]. Alam, T. M., et al. (2019). A Model for Early Prediction of Diabetes. Informatics in Medicine Unlocked, 16, 100204.
- [10]. Priyanka, B., and Singh, H. P. (2021). A Review on Big Data Analysis of Tobacco Consuming Trends in India. Annals of the Romanian Society for Cell Biology, 25(4), 3516-3527.
- [11]. Sharma, T. (2021). A Comprehensive Review of Machine Learning Techniques on Diabetes Detection. Visual Computing for Industry, Biomedicine, and Art, 4, 30.
- [12]. Naiyer, V., Sheetlani, J., and Singh, H. P. (2020). Software Quality Prediction Using Machine Learning Application. In Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2 (pp. XXX-XXX). Springer Singapore.
- [13]. Pasha, S. I., and Singh, H. P. (2021). A Novel Model Proposal Using Association Rule Based Data Mining Techniques for Indian Stock Market Analysis. Annals of the Romanian Society for Cell Biology, 25(6), 9394-9399.
- [14]. Rasool, M. A., Singh, H. P., and Reddy, K. N. (2021). Data Mining Approaches to Identify Spontaneous Homeopathic Syndrome Treatment. Annals of the Romanian Society for Cell Biology, 25(9), 3275-3286.